An Intonation Model for Embedded Devices Based on Natural F0 Samples

Gerasimos Xydas and Georgios Kouroupetroglou

Department of Informatics and Telecommunication
University of Athens
{gxydas, koupe}@di.uoa.gr

Abstract

The evolution of hand-held devices has made possible the porting of high quality Text-to-Speech systems to embedded platforms. However, linguistic resources required to build natural-sounding prosody models still need to be scaled down, to meet the hardware specifications of the devices. In this work, we present a compact intonation model that brings together the naturalness of corpus based prosody modeling with the limited nature of the embedded TtS applications. A sampling process of 3 points per syllable over a small set of appropriately set up utterances is used as the tonal unit database. The sampled points are applied at synthesis time over an onset/offset syllabic structuring of phrases. The model requires less than 1KB of storage for modeling each prosodic phrase class. The application to the Greek language is being demonstrated utilizing only lexical stress information.

1. Introduction

One of the most important tasks in Text-to-Speech (TtS) synthesis is the generation of the appropriate F0 contours of utterances. The overall quality of a TtS can be heavily affected by the quality of the underlying prosody model. Corpus-based prosody modelling can yield high quality natural sounding prosody for TtS synthesis [1]. However such models require a big amount of linguistic information to apply on, such as part-of-speech tagging, shallow syntax information, morphological structure, discourse information etc. The highest the target quality is, the most sophisticated the linguistic information required [2].

On the other hand, the evolution of embedded devices, such as cellular phones and PDAs, in terms of memory and CPU capabilities has made possible the porting of several TTS systems to these platforms. The expectations concerning the speech quality are not as high as in telecom applications, taking into account the hardware restrictions and the application domain of the small devices. Thus, degrades in several aspects of speech synthesis are reasonable. The major memory and CPU consuming linguistic resources that need to be degraded in embedded applications are the lexicons and the unit databases (e.g. diphones). Lately, a significant research focus has turned to the provision of TtS systems with small footprints [3][4].

Regarding corpus-based prosody modeling, the most common F0 contour generation in the public domain FESTVOX series [5][3], uses the Linear Regression (LR) method [6] to build three training models per syllable for F0 targets: one for the beginning of a syllable, one for the middle part of the vowel and one for the end of any voiced segment. This approach requires a big amount of training data, incorporating information of an enriched set of features like prosodic phrase breaks, part-of-speech, syllabic distances from major breaks, shallow syntax information etc.

In this work, we present a corpus-based model for generating F0 contours, by using samples from natural curves aligned on the above mentioned points over sequences of syllables. This differs from the LR approach in terms that it utilizes a database of F0 points, sampled from natural, human F0 curves. In our case, we do not build decision trees to predict these 3 pitch points. In contrast, the sampling process copies them from natural speech, annotated by syllabic structure information and stores them in the tonal unit database. During synthesis time, we select them by matching sequences of syllables with common features. In following sections, we focus on the minimum configuration of the model we propose to be used in embedded systems. Finally, we present a prototype for the Greek language. A database of 58 syllabic patterns for each prosodic phrase (PP) class was developed for the evaluation.

2. The F0 sampling model

According the Heterogeneous Relation Graphs [8] of FLITE [3], our tonal database utilizes only the Syllable and the Target relations of the recorded Utterances (Figure 1). Figure 2 illustrates the sampling process over a natural Utterance.

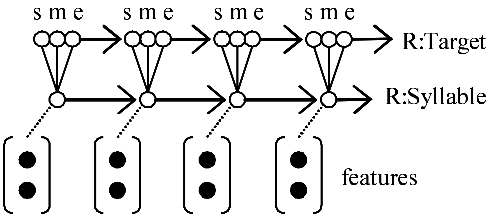

Figure 1: The Syllable relation together with the sampled Target one. (s=start, m=middle and e=end).

The F0 sampling model relies on the syllabic structure of a phrase. This forms the basis of the selection mechanism. Among the syllables, we identify those that can contribute to the description of the prosodic structure.

2.1. Modeling of syllabic patterns and phrases

The first principle of this light model is the de-compilation of the tonal events to a stack of layers. The lower layer, which is examined here, focuses to the description of a neutral F0 contour. A neutral model should produce reasonable-sounding results in any domain. Any focus, emphasis or de-emphasis information can be added on top of this layer as an

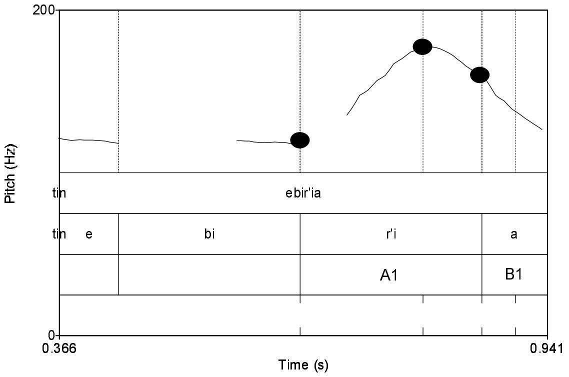

Figure 2: Sampling of pitch points over the stressed syllable “r’i” in the word “ebir’ia” *(experience).*

extension to the feature set used in the selection algorithm and is not discussed here.

An important factor to the success of this model was the outcome of several experiments that pointed out that the most perceivable portions of a neutral phrase is its *onset* and its *offset*, i.e. how does a phrase start and how does it end. Thus, we do not target to cover all possible sequences by decision trees but to efficiently encode the samples that belong to these sections with strong tonal interest. To achieve this, our approach takes into account a set of words at the beginning of a PP, called *onset,* and a set of words at the end of a PP, called *offset*. Samples encoding covers these sections. The middle section is left uncovered. During synthesis, we apply an exponential interpolation over the tonal parameters followed by simple “hat” patterns over accented syllables. This forms the base layer of the model. In order to keep the database to a minimum, we seek to model short onset and offset sections.

2.2. The onset and the offset

The phrase onset is concerned to be the portion from the first syllable of the phrase until the first accented word of it. For example in the phrase “το όνομα μου είναι Νίκος” (my name is Nikos) the onset is “το όνομα” (my name). Again, neutrality is something that needs further investigation, however we assume here that there is no any strong focus point to any paradigm.

On the other hand, the offset of a phrase needs more fine handling, as it characterizes boundary rises and falls in the F0 curve and so has bigger perceivable weight. Thus, the offset consists of the last two accented words of a phrase. In the above example, the offset is “είναι Νίκος” (is Nikos). As a result, in the Syllable relation, these sets are defined as:

- Onset = {H, A1, B1}
- Offset = {A2, B2, A3, B3, T}

where:

- **H(EAD)**: the very first syllable
- **A1**: the accented syllable in the first word of interest
- **B1**: the last syllable (boundary) in the word of A1
- **A2**: the accented syllable in the pre-last word of interest
- **B2**: the last syllable in the word of A2
- **A3**: the accented syllable in the last word of interest
- **B3**: the last syllable in the word of A3
- **T(AIL)**: the very last syllable

This is not a mandatory set but an optimum one. Other designs might extend these ranges, but the analysis corpus should grow as well. In any case, these are thought to be the syllables of *tonal interest.* It is obvious that we provide finest manipulation of the offset than the onset and this sounds true in real neutral speech.

The above eight (8) syllables are the maximum number. PPs can consist even of just one (1), depending on the amount of accented words and the possibility some of them to refer to the same syllable (e.g. “yes?” → A1=B1=HEAD). Based on that, the number of all the possible syllabic patterns is 58, some of which are tabulated in Table 1.

Table 1. Samples of syllabic patterns for PPs.

H	A1 T	H A1	H A1 B1 T
H T	A1 B1	H A1 T	A1 A3
A1	A1 B1 T	H A1 B1	A1 A3 T
…	…	…	…

2.3. Building the tonal database

The tonal database consists of recorded utterances with the aforementioned 3 sampled pitch points and syllabic annotation. In order for the syllabic patterns to be flexible to manipulations, sentences with long, accent-free middle parts are preferred for the recordings. The speaker should produce clear accent tones only in the onset and offset of the phrases, while the whole utterances should read out without any emphasis even in cases where emphasis is implied. In practice, such confusing prompts should be avoided. As a result, we manage to have high F0 resolution in cases of short phrases. In other cases, the curve loss in the middle section is kept to a minimum by avoiding having pitch accents in the there, so that accents can be appended during runtime on top of neutral contours.

2.4. Memory requirements

The 58 syllabic patterns result in the encoding of 263 syllables for each PP class. For each syllable we sample 3 pitch points (start, mid, end) that can be represented by one byte (0-255 Hz). That is a reasonable range for a male or female (in that case an overall shift can be applied). Thus, for each PP class the memory requirements are:

263 syllables * 1 byte * 3 point = 789 bytes

However, this low memory specification has some negative aspects, and that is what we tried to evaluate, because in long phrases, the middle section is not efficiently rendered.

3. Applied in Greek

The implementation of this model to the Greek language required the definition of the onset and the offset syllable sequences as well as the identification of the accented ones. To keep to a minimum, we did not lie on part-of-speech or any other information that requires a huge lexicon (more than 3G for Greek) as usual [2], but only on lexical stress, which is written in the case of Greek. The assumption we did was that in neutral speaking, stressed words get the accents. This is a rough approach that lies on the fact that most of the function words in Greek are unstressed and the results from the experiments validated this assumption. Furthermore, for each syllabic pattern, we model more than one PP class in order to include rising, falling and other tonal boundaries (Table 2).

Table 2. Examples of S(yllabic).PATTERNs and P(rosodic)P(hrase).CLASSes.

PHRASE	S.PATTERN	PP.CLASS

"τι ώ-ρα εί-ναι;"
(what time is it?)
	H(τι) A1(ώ)
	B1(ρα) A3(εί)
	B3(ναι)
		rising

"το πρό-βλη-μα της μό-λυ-νσης του πε-ρι-βά-λλο-ντος θα με-λε-τη-θεί στην συ-νά-ντη-ση αυ-τή."
(the problem of the pollution of the environment will be discussed at the meeting.)
	H(το) A1(πρό)
	B1(μα) A2(νά)
	B2(ση) A3(τή)
		falling

The model requires 58 syllabic patterns in the tonal database for every PP class. In the small-footprint cases, PPs are identified by spotting the punctuation marks of sentences. That is a simplified approach. Thus, 58 patterns are required for phrases ending to (a) a full stop (assigned the *falling* class), (b) a comma and (assigned to the *fall-rising* class), (c) a question mark (assigned to the *falling* class) etc. The selection algorithm chooses one out of these 58 alternatives based only on the syllabic structure and the identified class. Our Greek implementation includes (a), (b) and (c) class in a total of 174 utterances. These classes turned to be adequate for the nature of the short messaging services provided to the embedded devices.

3.1. Example

Here is how the Greek model works for the synthesized utterance:

"Το πρόβλημα της μόλυνσης του περιβάλλοντος πρόκειται να συζητηθεί, μεταξύ άλλων, στη συνάντηση αυτή."
(The problem of the environmental pollution is about to be discussed, among others, at the meeting)

In this example, the phrasing module classifies the given sentence into 3 PPs, as tabulated in Table 1. Figures 3 to 5 depict the selected tonal units from the database according to the S.PATTERN sequence.

Table 1. Phrasing table.

#	S.PATTERN	PP.CLASS
1	H A1 B1 A2 B2 A3	fall-rising
2	H A1 A3 B3	fall-rising
3	H A1 B1 A3	falling

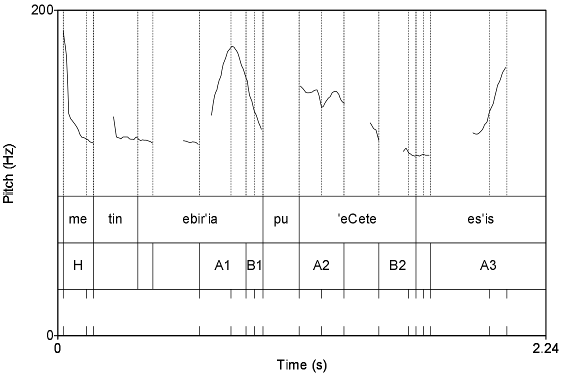

Figure 3: The tonal unit of "H A1 B1 A2 B2 A3" – fall-rising (1). The lower tier shows the 3 sampled pitch points stored per each syllable of interest.

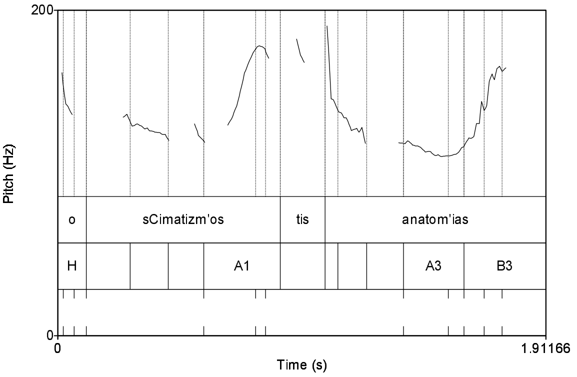

Figure 4: The tonal unit of "H A1 A3 B3" – fall-rising (2)

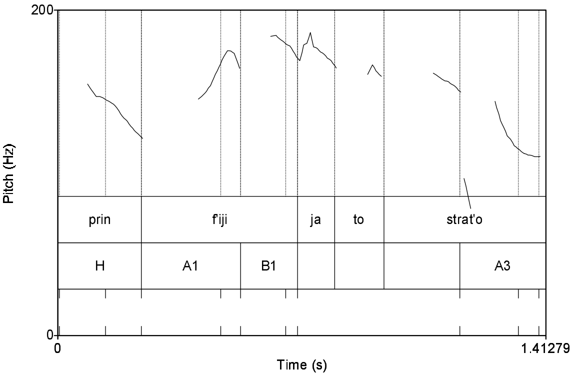

Figure 5: The tonal unit of "H A1 B1 A3" – falling (3)

Figure 6 illustrates the resulting synthesized F0. The tonal units that have been selected from the database are: 1:N has been rendered from unit 1 onset (sampled pitch points {H, A1, B1}), 1:F from unit 1 offset {A2, B2, A3}, 2:N from unit 2 onset {H, A1}, 2:F from unit 2 offset {A3, B3}, 3:N from unit 3 onset {H, A1, B1} and 3:F from unit 3 offset {A3}. The section 1:M was rendered by applying "hat" patterns over an exponential interpolation between 1:N and 1:F. This is not optimum for long middle sections; however it performs well in cases where 2-3 accented syllables are in there.

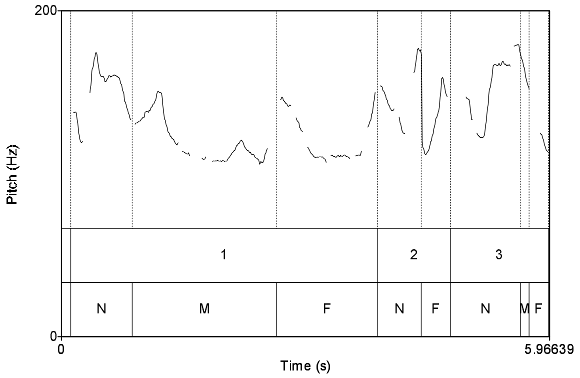

Figure 6: The synthesized utterance of the example. N=onset, M=middle and F=offset, while 1, 2 and 3 stands for the 3 selected tonal units.

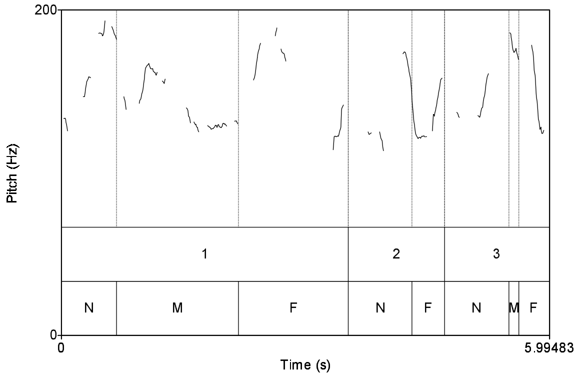

Figure 7: Natural intonation for the example.

Figure 7 shows the F0 curve from a human recording of the example's utterance. If you compare with the one presented in Figure 6 you can see the quality of the F0 peaks achieved by the 3-point sampling modeling. The samples of the above example can be found in http://www.di.uoa.gr/~gxydas/f0_sampling.

4. Discussion

The experiments we have done using the above model showed that this works well when the phrases are short, i.e. when the middle section is short, as we almost copy the F0 curve from the database, aligning it properly over the 3 pitch points per syllable. When the middle section extends to more than 2-3 accents, it sounds as monotonous as the exponential interpolation implies and this requires further improvements. However, it is a fact that most messages in mobile devices are quite short, such as SMS messages, appointments, address book etc. We have presented here the minimum configuration of the model using only the syllabic structure. Content/function word distinction, POS information or other linguistic clues can further improve the identification of the words of tonal interest that describe the onset and the offset of prosodic phrases, which ranges can be increased to provide higher resolution during modeling.

A degraded rule-based rendering version of the Greek framework presented here has been integrated in the free Greek TtS DEMOSTHeNES [7] that implements the same 58-patterned structuring but includes hand-written pitch points in the tonal database. The corpus-based one introduced by the F0 sampling method here delivers more natural-sounding prosody with high-resolution contours in the onset and offset sections (samples are available in the above mentioned URL). The complete proposed model has been integrated in a Greek version of FLITE (http://www.di.uoa.gr/~gxydas/greekflite).

5. Conclusions

We presented a novel intonation and F0 contour generation model that is based on 3 natural F0 samples per syllable from a tonal unit database. The advantages are that we can apply almost corpus-based prosody selection over a minimum of resources that can fit to the small footprints requirements of embedded devices. Furthermore, the proposed model is flexible enough to be extended when more enriched linguistic information is available. Finally, due to the nature of the samples, it is obvious that pitch points can be further stylized in order to achieve smoother transitions among heterogeneous concatenated contours. An implementation in the Greek language based on modeling lexical stresses produced natural sounding intonation compared to the rule-based model one.

6. Acknowledgements

The work of this paper has been partially supported by the GR-PROSODY project of the KAPODISTRIAS Program of the Special Account for Research Grants, National and Kapodistrian University of Athens.

7. References

[1] Taylor P. and Black A. W., "Speech Synthesis by Phonological Structure Matching", in Proceedings of Eurospeech'99, pp. 623-626, 1999

[2] Xydas G., Spiliotopoulos D. and Kouroupetroglou G., "Modeling Prosodic Structures in Linguistically Enriched Environments", 7th International Conference on Text, Speech and Dialogue, TSD2004, (to appear), 2004

[3] Alan W Black and Kevin A. Lenzo, "Flite: a small fast run-time synthesis engine", In Proceedings of the 4th ISCA Workshop on Speech Synthesis, pp. 204-207, 2001

[4] Tomokiyo M. L., Black W. A. and Lenzo A. K., "Arabic in my Hand: Small-footprint Synthesis of Egyptian Arabic", in Proceedings of Eurospeech'03, pp. 2049-2052, 2003

[5] Black A. W., Taylor P. and Caley R., "The FESTIVAL Speech Synthesis System", http://www.festvox.org, 1998

[6] Black A. W. and Hunt A., "Generating F0 Contours from ToBI Labels using Linear Regression", in Proceedings of ICSLP'96, 1996

[7] Xydas G. and Kouroupetroglou G., "The DEMOSTHeNES Speech Composer", in Proceedings of 4th ISCA Workshop on Speech Synthesis, pp. 167-172, 2001

[8] Taylor, P., Black, A., and Caley, R.: Heterogeneous Relation Graphs as a Mechanism for Representing Linguistic Information. Speech Communications 33, pp. 153-174, 2001